ACADEMIC PRESS

Available online at www.sciencedirect.com

SCIENCE DIRECT®

Journal of Biomedical Informatics 35 (2002) 222–235

Journal of Biomedical Informatics

www.elsevier.com/locate/yjbin

Two biomedical sublanguages: a description based on the theories of Zellig Harris

Carol Friedman,[a,b,*] Pauline Kra,[a] and Andrey Rzhetsky[a,c]

[a] Department of Medical Informatics, Columbia University, VC5, Vanderbilt Building, 622 West 168th Street, New York, NY 10032-3720, USA
[b] Department of Computer Science, Queens College CUNY, 65-30 Kissens Blvd., Flushing, NY 11367, USA
[c] Genome Center, Columbia University, 1150 St. Nicholas Blvd., Russ Berrie Pavilion, Informatics, Room 121H, New York, NY 10032, USA

Received 15 August 2002

Abstract

Natural language processing (NLP) systems have been developed to provide access to the tremendous body of data and knowledge that is available in the biomedical domain in the form of natural language text. These NLP systems are valuable because they can encode and amass the information in the text so that it can be used by other automated processes to improve patient care and our understanding of disease processes and treatments. Zellig Harris proposed a theory of sublanguage that laid the foundation for natural language processing in specialized domains. He hypothesized that the informational content and structure form a specialized language that can be delineated in the form of a sublanguage grammar. The grammar can then be used by a language processor to capture and encode the salient information and relations in text. In this paper, we briefly summarize his language and sublanguage theories. In addition, we summarize our prior research, which is associated with the sublanguage grammars we developed for two different biomedical domains. These grammars illustrate how Harris' theories provide a basis for the development of language processing systems in the biomedical domain. The two domains and their associated sublanguages discussed are: the clinical domain, where the text consists of patient reports, and the biomolecular domain, where the text consists of complete journal articles.

1. Introduction

With the constantly increasing availability of online textual information and computational power, we experience increased utilization of natural language processing (NLP) techniques in the biomedical domain. In the 2001 and 2002 AMIA Fall Symposiums there were 10 and 14 papers, respectively, on natural language processing as compared to a handful of papers in sessions before 1998. Similarly, the Pacific Symposium on Biocomputing (PSB) has experienced an increased interest in the use of NLP for mining the literature for knowledge acquisition and improving retrieval of the literature. The NLP sessions at the 2001, 2002, and 2003 PSB conferences attracted 10, 17, and 15 submissions, respectively, from which 4, 6, and 5, respectively, were accepted. In addition, in 2002 and 2003 the Association of Computational Linguistics held the first two workshops on biomedical language processing.

NLP methodology has been used to obtain and structure clinical and biomolecular information. Diverse NLP clinical applications have been developed to be used for decision support [1,2], encoding [3–6], data mining and clinical research [7–9], order entry [10], information retrieval [11–13], and controlled vocabulary [14]. A number of evaluations of these applications demonstrated they were effective for realistic clinical applications. Additionally, NLP has been used to improve access to the biomedical literature. In the last few years, a substantial effort has been associated with identifying biomolecular substances. One type of system primarily identifies gene or protein names in biological texts, [15–18], and others extract relations between the substances in order to automatically acquire knowledge and to improve retrieval of information [19–26].

Zellig Harris proposed a theory of sublanguages [27,28] that explains why it is possible to process language in specialized textual domains, such as those

[*] Corresponding author. Fax: 1-212-305-3302.
E-mail address: friedman@dmi.columbia.edu (C. Friedman).

1532-0464/02/$ - see front matter © 2003 Elsevier Science (USA). All rights reserved.
doi:10.1016/S1532-0464(03)00012-1

found in genomics and medicine. According to Harris, the languages of technical domains have a structure and regularity, which can be observed by examining the corpora of the domains, and which can be delineated so that the structure can be specified in a form suitable for computation. Whereas the general English grammar theory primarily specifies well-formed syntactic structures only, Harris's sublanguage grammar theory also incorporates domain-specific semantic information and relationships to delineate a language that is more informative than English because it reflects the subject matter and relations of a domain as well as the syntactic structure.

The scientific grounding of Harris's sublanguage theory is well established and has been repeatedly verified by the vast amount of work that has been done in this area. A set of papers on sublanguage processing and research collected by Grishman and Kittridge [29] includes the domains of lipoprotein kinetics [29] clinical patient reports [30], telegraphic navy messages [31,32], and reporting of events in outer space [33]. Additional work pertaining to the sublanguages of pharmacological literature and lipid metabolism is described in Sager [34]. An earlier collection of papers on this subject was edited by Kittredge and Lehrberger [35]. We find that Harris's principles are applicable to much of our work in biomedical language processing. In this paper we describe properties of the languages of two specialized domains in biomedicine, the clinical domain and the biomolecular domain, and show how Harris provides a linguistic foundation for our work.

In the next section we provide background material by summarizing important aspects of Harris's principles on language and sublanguage. We then present the first major use of those principles, which resulted in the development of a comprehensive sublanguage grammar and medical language processing system by Sager [36] who heads the Linguistic String Project. The subsequent section analyzes features of the language of the clinical domain, which is followed by a section on the language of the biomolecular domain. Finally, we discuss differences and similarities of the two sublanguages, and provide our conclusion.

2. Background

2.1. An overview of Harris' theory on language and sublanguage

Harris postulated that all occurrences of language are word sequences satisfying certain constraints which express and transmit information. The constraints are: dependency relations [28; 54–61], paraphrastic reductions [28; 79–96], and inequalities of likelihood [28; 61–79]. Additionally, certain subsets of languages exist (i.e., sublanguages) within specialized domains that exhibit specialized constraints due to limitations of the words and relations of the subject matter [28; 272–318].

Dependency relations are concerned with syntactic regularities, and are applicable to all general language as well as to specialized languages. The crucial property is the dependence of words in a sentence on other words, and the categorization of words accordingly. Basically, words that are nouns or concrete objects are considered zero-level words (e.g., *cats*, *fish*) because they do not depend on other words in the sentence. More specifically, zero-level words occur with other words that are first or second level words. In contrast, words that are verbs are either first-level or second-level words, which are considered to be operators that are dependent on their arguments. Sentences containing a first-level word (i.e., *eat* as in *cats eat fish*) must contain zero-level words that are arguments of the first-level words. Likewise second-level words (e.g., *knows*) have at least one argument that is a first-level word. For example, in *John knows cats eat fish*, the object argument of *knows* is a sentence containing a first-level operator *eats*, which has two zero-level arguments *cats* and *fish*. This language component is concerned with classes of words but not with individual words, and permits strange or unlikely combinations, such as *computers eat fish*, as long as the dependency constraints are met.

Inequalities of likelihood. The dependency relations exhibit different likelihood constraints. For example, certain arguments are more likely to occur with certain operators than with others. Thus, *cat* is more likely to occur as the first argument of *eat* than *table* or *concept* is. The likelihoods of an operator in respect to particular arguments are based on the frequency of operator-argument combinations; some combinations occur frequently whereas others occur very rarely. In general language, the likelihood constraints on operators and their arguments are fuzzy, while in sublanguages the constraints are generally sharper. In either case, combinations that have a very high likelihood create a low information situation, in which case zeroing of highly likely words may occur. For example, in general language, the indefinite noun *something* is often zeroed, as in *I ate* and *I read*, without loss of information. In the clinical language there is no loss of information if the noun *patient* and verb *has* is zeroed, as in *fever*.

Paraphrastic reductions involve transformations of the sentences from a simple primitive form (which we typically do not observe in text) to a complex form that consists of the actual sentences we see in textual documents (i.e., called the surface form). These reductions are paraphrastic in that they are associated with operations that change the structure of the sentence without changing the informational content. These reductions serve the purpose of eliminating information that is highly likely to occur and that is redundant by shortening and combining the original sentences so that they represent a more efficient and compact form. For example, the sentence *John has a brown hat*, is reduced from a single sentence *John has a hat*; *the hat is brown*. When looking at natural language text, we do not see the original sentences because they generally have undergone numerous transformations to become the surface forms that constitute the sentences we do see in text.

Sublanguage grammar. When we look at sentences within a specialized domain, it is possible to observe particular word classes and particular statement types where the statement types generally contain operators that are much more restrictive than the dependency relations permitted in the general language and the likelihood constraints are much more definitive. For example, in a general language it is permissible, although not sensible, to say *John activated protein A*, because the syntactic combination of word classes is well-formed. However, in the biomolecular language domain, this sentence is not legitimate because the operator *activate* permits only certain combinations of the word classes (i.e., *substance activate substance*, *process activate substance* are allowed but *person activate substance* is not), and the sublanguage operators reflect the salient relations and arguments that are meaningful in the specialized domain.

Specialized sublanguages deal with specific subject matter (i.e., textual radiological reports, discharge summaries or other patient documents, biomolecular literature, medical literature, etc.). In this paper, we focus on two sublanguages within the biomedical domain: clinical reports and biomolecular relations found in the literature. The specifications of these sublanguages resulted in the implementation of two sublanguage grammars, which were used to process text and to extract and structure relevant information. It was possible to create these grammars because of the regularities and co-occurrence relations within each of the specialized sublanguages. Immunology literature is also in the biomedical domain, and Harris discussed the immunology sublanguage as presented in a companion article in the same issue of this journal. A more extensive analysis of the immunology sublanguage is discussed by Harris et al. [37].

In the grammar of a specialized sublanguage, operators and arguments still satisfy the dependency relations of the whole language and paraphrastic reductions still occur, but the vocabulary is limited, only restricted combinations of words occur, and subclasses of words combine in specified ways with other subclasses. In a sublanguage, words form subsets from the larger word classes of the overall language. In the biomolecular domain, subsets of classes can be identified which correspond to words denoting proteins, genes, cells, tissues, and other biomolecular substances that constitute the relevant objects corresponding to the subject matter of the domain. Moreover, words that do not belong to relevant word classes of the domain (e.g., *pencil*, *desk*) are excluded from the sublanguage. We can also form subsets from other classes, such as the verb class, that depend on the subclasses of the arguments that co-occur with the verbs. For example, the combination *Fyn activates Cbl* is a well-formed pattern because the sequence PROTEIN ACTIVATE PROTEIN occurs regularly, but the combination *liver cells bind to protein B* (CELL ATTACH PROTEIN) is not allowed since that particular combination of word classes never occurs.

Thus, in order to create a sublanguage grammar, the critical task is to discover the subclasses and important relations. For each domain, clustering techniques [38] help to discover a limited number of word classes and sentence types for a large sample of a domain corpus. However, the sentences are in surface forms, and therefore, many reductions have occurred so that the sentences are complex and not necessarily in forms close to the underlying operator-argument forms, making the discovery task more difficult.

There are several other elements of Harris's theory concerning sublanguages. A sublanguage may differ from the whole language by omitting some grammatical properties of the language or by allowing different reductions. The domain-specific syntactic constraints and reductions are not necessarily the same as those of general English. We have observed this feature within the clinical domain because many well-formed sentences are telegraphic, in that they often are missing subjects and verbs, because that information is implicit in the context. For example, in a section of a report corresponding to **chief complaints**, sentences may consist of noun phrases only. In this context a noun phrase, such as *pain*, has an operator, such as *is associated with the patient*, which is expected in this context and therefore can be omitted.

Another interesting observation is that sublanguages may overlap because they are associated with some of the same events or entities. For instance, the clinical and biomolecular domains are concerned with tissues and diseases, but the underlying relationships associated with them differ substantially. That may imply that similarities and differences in sublanguages or overlapping scientific fields may be quantifiable by measuring differences and similarities in words classes and their membership in co-occurrence patterns.

3. Sublanguage features

In the following, we discuss features of languages in specialized domains that have important implications for the development of computerized natural language processing systems:

- *Semantic categorization of words*. Relevant words can be categorized into subclasses or types of information where the types form the underlying subject matter of the domain. For example in the clinical domain there are informational categories, such as: disease, procedure, body location, and medication, whereas in the biomolecular domain some relevant categories are: gene, protein, amino acid, small molecules, and interaction.

- *Co-occurrence patterns and constraints*. Particular subclasses of information combine in particular co-occurrence patterns to form the meaningful relations of the domain. In the clinical domain a relation such as, PATIENT $\text{VERB}_{\text{HAVE}}$ SYMPTOM IN BODYLOCATION (e.g., *patient experienced pain in joints*) specifies that a patient is experiencing a symptom in a body location. In the biomolecular domain, a relation, such as PROTEIN_1 $\text{VERB}_{\text{INTERACTION}}$ PROTEIN_2 (e.g., *Fyn activated Cbl*), would be appropriate, whereas the former pattern most likely would not be. In addition, it is possible to refine a class, such as INTERACTION, to further constrain the arguments it combines with. For example, if we create two subclasses of INTERACTION, called ACTIVATE and ATTACH, then we can refine the well-formed patterns so that PROTEIN $\text{VERB}_{\text{ACTIVATE}}$ PROCESS (e.g., *Akt activates apoptosis*) would be considered a valid pattern, whereas PROTEIN $\text{VERB}_{\text{ATTACH}}$ PROCESS (e.g., *Cbl binds to apoptosis*) would not be.
- *Paraphrastic patterns*. A set of patterns represent an equivalence relation where the patterns are different grammatically but represent the same underlying operator-argument structure. Thus, in the clinical domain, the patterns BODYLOCATION VERB_{BE} SYMPTOM (e.g., *joints were painful*), SYMPTOM IN BODYLOCATION (e.g., *pain in joints*), SYMPTOM BODYLOCATION (e.g., *painful joints*), and BODYLOCATION SYMPTOM (e.g., *joint pain*) all are equivalent to PERSON $\text{VERB}_{\text{HAVE}}$ SYMPTOM IN BODYLOCATION. Similarly, in the biomolecular domain, the patterns PROTEIN_2 VERB_{BE} $\text{VERB}_{\text{INTERACTION}}$ BY PROTEIN_1 (e.g., *Cbl was activated by Fyn*), and PROTEIN_2 $\text{NOUN}_{\text{INTERACTION}}$ OF PROTEIN_1 (e.g., *Fyn activation of Cbl*), are equivalent to PROTEIN_1 $\text{VERB}_{\text{INTERACTION}}$ PROTEIN_2 (e.g., *Fyn activated Cbl*). It is therefore possible to choose one of the patterns as representative of the type of informational relationships conveyed by the set.
- *Omission of information*. In a specific domain, when the context is known, additional contextual information is often expected or understood. According to Harris, information that is expected is low in information content and can be omitted because it is recoverable from the context. For example, in a radiological report of the chest, *infiltrate noted* is interpreted to mean *infiltrate in lung was noted by radiologist*. Omitted information is troublesome for language processing because a system must have additional knowledge in order to recover all the implicit information.
- *Intermingling of sublanguage patterns and general language*. When looking at text of a domain, the sublanguage patterns are often interspersed with general language that is not in the sublanguage, making the process of identifying sublanguage co-occurrence patterns difficult. For example, in *he complained of a severe headache while working in the family store*, *and then fainted*, the expression *while working in the family store* may be relevant to the patient's condition but is not in the clinical sublanguage because it does not contain the sublanguage entities or relations while the rest of the sentence does.
- *Terminology*. Within specialized domains, words often take on different meanings than in the general world domain, and therefore specialized domain vocabularies are needed to process domain text. For example, in the clinical domain *capsule* may denote a body location component, and in the biomolecular domain, *associates* may denote an interaction sense *binds* in addition to its general English sense. In multi-word terms the issue is more complex because a term may have a meaning that is compositional and therefore denotes the meaning of the components, or that is non-compositional and denotes something different than the combined meaning of the components. For example, in radiological reports of the chest, *no active disease* not only means that there is no current disease activity but also denotes that there are signs of a previous condition in the X-ray. Thus, the phrase *no active disease* takes on a specialized meaning in the domain as if it were a single word. Evidence supporting this assumption is the frequent occurrence of the phrase in radiological reports and the existence of a corresponding abbreviation *NAD*. The issue concerning compositionality is not a trivial one, and a number of articles have been published in the medical domain that discuss compositionality [39–41] and modification [42]. The main issue is whether or not to treat certain multi-word terms, such as *chest pain* and *chronic cough*, as single words, or as words with modification. This issue is even more complex in the biomolecular domain because (1) verbs play a more significant role, (2) verbs frequently occur in the nominal form (e.g., *activation*, *activator*), (3) verbal relations are often nested, and (4) substances are frequently given complex names that correspond to the functions they perform. Thus, *inhibitor of mitogen activation* may refer to the name of a particular entity or to a type of entity (i.e., one that inhibits mitogen). In the case of *mitogen-activated protein kinase* there is evidence to support considering the phrase as a single name: the phrase is frequently found in the literature, and has a corresponding acronym, MAPK. We believe that more research on this topic is needed. The use of large domain corpora and statistical analysis of frequency distributions may provide us with

objective criteria for determining whether or not a term should be treated as a single unit (called a collocation in the computational and statistical literature), or a compositional phrase composed of separate elements. Manning and Schütze [43] provide a discussion about linguistic and statistical considerations of collocations, which roughly are phrases that are not straightforward compositions of the parts.

- *Controlled vocabulary*. Controlled vocabularies are usually associated with domain-specific terminology, well-defined concepts, and with a method for linking the terms in the terminology to the concepts. These vocabularies were generally defined based on expert knowledge, and as such are not definable according Harris because they were not established on distributional grounds. A controlled vocabulary is important for computerized applications because it facilitates sharing of information among different systems by making it possible to map the multiple ways of expressing a concept to one well-defined concept in the controlled vocabulary, thereby reducing the variety of expression after the mapping. However the mapping between language and any controlled vocabulary is somewhat arbitrary and difficult to justify on objective grounds. In the clinical domain there are several controlled vocabularies, such as the UMLS [44], SNOMED [45], ICD-9 [46], and MESH [47], which consist of well-defined or standard concepts corresponding to words and phrases in the domain. In the biomolecular domain there are a number of systems that are associated with terminology and controlled vocabularies, such as the Gene Ontology [48], GenBank [49], SWISS-PROT [50], and LocusLink [51].
- *Ontology*. If applications using a controlled vocabulary require reasoning, a formal specification of the entities and relations in the domain (i.e., an ontology) is very useful. An ontology of a domain may overlap with semantic classes associated with a domain sublanguage. For example, the clinical domain and biomolecular domain are likely to have classes corresponding to body location, disease, and medication (classified as small molecules in the biomolecular domain). However, in an ontology, the classes are based on knowledge of the domain and are used to facilitate reasoning. In a sublanguage, classes are used to recognize, constrain, and interpret co-occurrence patterns. A big difference between the two is that sublanguage patterns are obtained through objective analysis of data, while ontologies are not necessarily constructed with direct empirical evidence. An ontology may be useful for natural language processing applications, but this is not necessarily true because the granularity of the classes may differ. For example, it is generally sufficient for a sublanguage to have one coarse class called DISEASE (or an even coarser class FINDING) without having a complex hierarchy of disease subclasses because all the members of the class DISEASE generally have the same co-occurrence patterns. However, in an ontology, it would be preferable to partition the class into subclasses since applications involving reasoning would generally treat the subclasses DISEASE, RESPIRATORY DISEASE, PULMONARY DISEASE, and PNEUMONIA differently because they are associated with different clinical properties.

4. The sublanguage of the clinical domain

4.1. Background

Harris provided a theoretical basis for sublanguage processing and identified semantic categories and co-occurrence patterns for several scientific languages. Two large scale natural language processing systems, the Linguistic String Project system [36] and the MedLEE system [52], were both based on Harris's theories and were both applied to the clinical domain. However, these systems differ from Harris because the grammars they each use follow a constituent grammar formalism and not an operator-argument formalism. Below we present a brief overview of the sublanguage approaches of the two systems.

The Linguistic String Project (LSP), headed by Sager et al. [36], was a pioneering effort in language processing in the general English domain and also in the clinical domain. A detailed version of a computerized English grammar and parsing system is described by Sager [53]. The LSP system was the first general comprehensive NLP system in the medical domain that contained a sublanguage grammar based on Harris's sublanguage theory. The LSP system inspired several other systems, including the early version of PROTEUS [29], PUNDIT [54], KERNEL [55], and MedLEE [52]. The LSP system established 40 clinical subclasses that denoted the relevant types of clinical information found in patient documents (i.e., symptom, medication, body part), and 14 general English semantic subclasses associated with verbs (e.g., *have*, *be*), temporal information (e.g., *change*, *increase*), evidential information (e.g., *no*, *present*), and connective operators (e.g., *consistent with*, *and*). In addition, six types of semantic relations were established that corresponded to the representation of basic informational relations associated with patient management (*transferred to ICU*), treatment excluding medication (*intubated*), medication (*penicillin qd*), tests and results (*ppd positive*), patient behavior (*drinks excessively*), and patient state (*pain in joints*).

In the LSP system, each statement type could be thought of as a frame or template representation that denoted predetermined semantic relations among word classes. For example, a statement type corresponding to patient state would be a template with slots for the patient state, temporal information, body location, severity, and evidence where the latter four types of slots represent optional qualifiers of the patient state (e.g., as in *patient experienced severe pain in joints yesterday*). Similarly, a medication statement type would have slots for medication, dose, frequency, manner, etc. (e.g., *on ampicillin 2 mg qid po*). Interestingly, the six statement types accounted for the majority of the relevant relations in the clinical domain.

MedLEE, which was developed by Friedman et al. [52,56] is also a comprehensive language processing system in the clinical domain that relies on a sublanguage grammar, which (i) specifies the subclasses in the language, (ii) delineates the structure of the language (i.e., well-formed sequences of subclasses), (iii) interprets the relations among the elements of the structures, and (iv) specifies a representational form for each structure, which is consistent with a formal representational schema for medical language [52]. The schema was designed on the basis of experience with the LSP system and a manual analysis of sample reports that were randomly retrieved from the clinical repository of patient reports at Columbia-Presbyterian Medical Center (CPMC).

MedLEE currently consists of 53 semantically relevant classes, most of which are shown in Table 1, and several syntactic classes that have semantic interpretations such as conjunction (e.g., *and*, *or*), preposition (e.g., *after*, *in*), and certain types of verbs (e.g., *involved*, *demonstrated*). The original schema was designed to represent findings in radiological reports, and was expressed in the format of conceptual graphs, but was later changed to a frame representation in the form of lists because it was more convenient computationally. According to the MedLEE schema the main relations in a radiology report consist of primary findings with optional modifiers (e.g., *moderate left posterior central gyral hypodensity* and connective relations between the primary findings (e.g., *CT scan* ***revealed*** *a hypodensity* ***consistent with*** *an infarct*). In the two examples the primary information is underlined and the connective relations are shown in bold. When expanded to broader domains such as discharge summaries, additional basic relations were added to represent new types of primary clinical events, such as medication, laboratory tests, demographic information, and behavior. Examples of the primary types of information and target forms are illustrated in Table 2.

Table 1
Semantic categories and subcategories in the clinical domain and examples

Primary category	Subcategory	Examples
ADT		*Admitted*, *transfer*
Behavior		*User*, *drinks*
Bodyfunc		*Breathing*, *movement*
Bodymeas		*Pulse*, *weight*
Device		*Catheter*, *atrial electronic pacemaker*
Finding		
	Cfinding	*Cardiomyopathy*, *diabetes mellitus*
	Descriptor	*Patchy*, *egg shaped*
	Organism	*E. coli*, *Staphylococcus*
	Pfinding	*Enlarged*, *opacity*
Labproc		*Liver function test*, *SMAC*
Labtest		*Sodium*, *alkaline phosphatase*
Med		*Aspirin*, *ace inhibitor*
Proc		*Biopsy*, *collapse therapy*
	Examproc	*X-ray*, *spectral doppler imaging*
Substance		*Cigarettes*, *illegal substance*
Technique		*Underpenetration*, *expired film*
Modifier category		
Bodyloc		*Heart*, *respiratory system*
Certainty		*Possible*, *rule out*
Change		*Increased*, *came down to*
Degree		*Slight*, *extensive amount*
Diagmaterial		*Barium*, *indium 131*
Ethnic		*Dominican*, *Hispanic*
Examphase		*Arterial phase*, *blood pool phase*
Examtype		*Serial*, *digital subtraction*
Family		*Mother*, *sister*
Frequency		*Bid*, *times two*
Locative		*Below*
Manner		*Intravenous*, *continuous infusion*
Position		*Axial*, *medial lateral oblique*
Ptactivity		*Awake*, *lying down*
Ptdescr		*Twin*, *left handed*
Quantity		*Multiple*, *one half*
Race		*Caucausian*, *black*
Reaction		*Respond*, *hypersensitive*
Region		*Left*, *right*, *upper*
Service		*Emergency room*, *ICU*
Sex		*Female*, *male*
Specialist		*Cardiologist*, *pathologist*
Speciality		*Cardiology*, *pathology*
Timeunit		*Day*, *morning*
Unit		*mg*, *centimeters squared*
Relational operator		
Conjunction		*And*, *or*, *as well as*, *with*
Connective		*Accompanying*, *including*, *consistent with*
Certainty verb phrase		*Appeared*, *cannot be excluded*

4.2. Clinical sublanguage

Table 1 lists the semantic categories in the clinical domain along with examples. Some of the classes correspond to primary types of information, such as BEHAVIOR, FINDING, and MEDICATION (e.g., *drinks*, *pain*, and *aspirin*). Additionally, some of the categories, such as FINDING, have been subdivided into subcategories

Table 2
Simplified co-occurrence patterns illustrated with examples and target representational forms

Category	Simplified patterns	Examples	Target form
Behavior	Substance + behavior Behavior + substance	*Cigarette smoker* *Smokes cigarettes*	[behavior,smoke, [substance, cigarettes]]
Bodyfunc	Bodyfunc + finding Finding + bodyfunc	*Walking with difficulty* *Difficulty walking*	[problem,difficult, [bodyfunc,walk]]
Device	Bodyloc + device	*Left ventricular assist device*	[device,assist device, [bodyloc,ventricle, [region,left]]]
Finding	Finding + in + bodyloc Bodyloc + v_{show} + finding	*Rash in arm* *Arm had a rash*	[finding,rash, [bodyloc,arm]]]
Finding	Change + Finding Finding + v_{change} Finding + v_{show} + change	*Increased pain* *Pain increased* *Pain did increase*	[finding,pain, [change,increase]]
Labtest	Weight + of + measure Labtest + v_{be} + measure	*Weight of 125 lbs* *Weight was 125 lbs*	[labtest,weight, [measure,[125,lb]]]
Proc	Proc + of + bodyloc Bodyloc + Proc	*Biopsy of breast* *Breast biopsy*	[proc,biopsy, [bodyloc,breast]]

CFINDING (complete finding involving the overall patient (*hypertension*, *diabetes*) or a specific body location (*cardiomyopathy*, *enlarged heart*), PFINDING (partial finding, such as *enlarged*, *tender*) where a location has not been specified, ORGANISM (*acinetobacter*, *clostridium difficile*), and SYMPTOM (*chills*, *fever*). The primary categories represent the basic informational entities in the sublanguage. Some are found to occur in sentences by themselves without any verbs or other relational operators. For example, in the chief complaint section of a discharge summary, a sentence *chills* can be found, whereas in the medications section a sentence *aspirin* can be found. Even though these types of relations each only consist of one word, and therefore do not explicitly contain operators as described by Harris, they are still consistent with Harris' theory. The reason that a sentence *chills* is reasonable is that readers assume that the operator is a verb, such as *had*, or *was given*, which was omitted because it is expected. In addition, in each case, it can also be assumed that an argument, *patient*, was omitted.

Categories of information that are not considered primary are modifier or secondary types of information. These types do not occur by themselves because they have meaning only when they modify other concepts or relations. Modifiers may modify primary types or may modify other modifiers. Examples of modifier classes are also shown in Table 1. Some are related to temporal information, such as onset (*intermittent*, *sudden*), evidential information (*rule out*, *no evidence*, *appears*, *has*), severity information (*mild*, *extensive*), body location type of information (*arm*, *left lower lobe*), and descriptive information (*patchy*, *amorphous*).

Table 2 shows some of the co-occurrence patterns for the domain and specifies their target form. One informational relation in the language system may be associated with multiple co-occurrence patterns, since there are often many different ways to express the same information. For example, the patterns:

- BODYLOCATION VERB$_{be}$ SYMPTOM (e.g., *joints were painful*),
- SYMPTOM IN BODYLOCATION (e.g., *pain in joints*),
- SYMPTOM BODYLOCATION (e.g., *painful joints*), and
- BODYLOCATION SYMPTOM (e.g., *joint pain*)

all have the same underlying structure as the pattern PATIENT VERB$_{have}$ SYMPTOM IN BODYLOCATION, are all associated with the same interpretation, and therefore are also associated with the same target form. The target forms are represented as frames in the form of lists where the first element in the list denotes the primary type of information, the second denotes the value, and the remaining elements are frames, which modify the primary type of information. For example, the target form for *cigarette smoker* is a frame denoting behavioral type of information *smokes*. It has a substance modifier, which has the value *cigarettes*.

Table 3 shows sample patterns for modifier categories and their interpretation. For example, the pattern DEGREE CHANGE is interpreted as a CHANGE type of modifier where the DEGREE type of information is operating on the CHANGE information. Modifiers occur frequently with primary categories and also with other modifiers. For example, the CHANGE modifier can also modify finding and procedure types of information, although this is not shown in the simplified modifier patterns shown in Table 3.

5. The sublanguage of the biomolecular domain

5.1. Background

A second comprehensive NLP system called GENIES was developed by Friedman et al. [57] for the biomo-

Table 3
Examples of modifier patterns for certain categories and the corresponding target forms

Category	Modifier pattern	Example	Target form
Bodyloc	Region + bodyloc	*Left arm*	[bodyloc,arm,[region,left]]
Bodyloc	Bodyloc + bodyloc	*Facial hair*	[bodyloc,hair,[bodyloc,face]]
Certainty	Negation + certainty	*No evidence*	[certainty,no]
Certainty	Degree + certainty	*Slight possibility*	[certainty,possible,[degree,slight]]
Change	Degree + change	*Slight change*	[change,change,[degree,slight]]
Degree	Degree + degree	*Very severe*	[degree,severe,[degree,very]]
Time	Quantity + Timeunit	*2 weeks*	[timeunit,[2,week]]

lecular domain, which was implemented by using the MedLEE system with a different sublanguage grammar. A sublanguage grammar specific to the biomolecular domain was developed in order to extract and structure biomolecular interactions from the literature that are associated with signal transduction and biochemical pathways within complex multicellular organisms as well as yeast and bacteria. Domain expertise was used to determine the important objects, entities, properties, and events in the domain, and an ontology was established by Rzhetsky et al. [58]. The ontology was based on expert knowledge and on manual analysis of the information in 300 online journal articles in *Science*, *Nature*, *Cell*, *Proceedings of the National Academy of Sciences of*

Table 4
Semantic categories and subcategories in the biomolecular domain

Categories	Subcategories	Examples
Substance		
	Geneorprotein	*pip2*, *c-myc*
	Gene	*il-2*, *p53*, *Caveolin-1*
	Protein	*Cbl*, *Fyn*, *Let-23*, *Myc-p70s6kE389D3E*
	Aminoacid	*Tyrosine*, *threonine 229*
	Small molecule	*zDEVD*, *guanosine triphosphate*, *tetracycline*
	Domain	*Src homology 2*
	DNA	*DNA*
	RNA	*mRNA snRNA*
	DNA region	*Origin of replication*, *codon 249*
	Cell	*Jurkat cell*, *LBL-DR7 cells*
	Structure	*Polyvinylidene difluoride membrane*
	Tissue	*Adrenal glomerulosa*, *astrocytoma*
	Species	*Human*, *Epstein–Barr virus*
Action		
	Activate	*Activate*, *induce*, *mediate*, *stimulate*
	Inactivate	*Inhibit*, *suppress*, *block*, *arrest*
	Attach	*Bind*, *join*, *immunoprecipitate*
	Actupon	*Affect*
	Breakbond	*Cleave*, *demethylate*, *proteolyze*
	Contain	*Contain*, *include*
	Cause	*Result in*, *lead to*
	Createbond	*Phosphorylate*, *polymerize*
	Express	*Express*, *overexpress*
	Generate	*Produce*
	Modify	*Mutate*
	Promote	*Catalyze*, *medate*, *enhance*
	React	*React*, *interact*
	Signal	*Up regulate*, *control*, *modulate*
	Substitute	*Replace*, *substitute*
	Transcribe	*Transcribe*
Process		
	Process	*Apoptosis*, *survival*, *mitosis*
	Pathway	*Ras pathway*
	Disease	*FAD*
State		*Mutant*, *active*, *anergy*
Relation		*And*, *or*, *homology*, *orthology*

the USA, and *Current Biology*, which were associated with the regulation of cell death in animals. The journal articles contained a vast amount of different types of information, but the domain expert determined which was the most relevant for capturing pathway information, and the sublanguage grammar was developed based on that ontology. This method of formulating the sublanguage classes and relations was a departure from Harris because it was based on use of external expertise rather than on distributional methods found in the text.

5.2. Biomolecular sublanguage

The sublanguage of molecular interactions is characterized by sentences that have high informational and structural complexity. The main type of information currently being captured by GENIES involves interactions, which are primarily expressed as verbal relationships or their corresponding nominal form. Additionally, the interactions are often highly inter-related because they not only interact with substances but also interact with other interactions, and therefore are nested. Because verbs are central to this domain and different inflected verb forms that correspond to actions (e.g., *inhibit*, *inhibition*, *inhibitor*, *inhibiting*, *inhibited*) are associated with different patterns, the sublanguage patterns include a combination of semantic and syntactic categories that also contain syntactic verbal subclasses. In particular, the verb subclasses that are included in the sublanguage patterns are: ACTION$\mathbf{v_p}$ (action verb, present tense—*inhibit*), ACTION$\mathbf{v_{ed}}$ (action verb, past tense—*inhibited*), ACTION$\mathbf{v_{en}}$ (action verb, past participle—*inhibited*), ACTION$\mathbf{v_{ing}}$ (action verb, progressive—*inhibiting*), ACTION$_\mathbf{n}$ (nominal form of action verb—*inhibition*), and ACTION$\mathbf{v_{or}}$ (agentive form of action verb—*inhibitor*).

Table 4 shows the sublanguage categories of the biomolecular domain with examples. Notice there are only five high level semantic categories specifically for this domain: SUBSTANCE, ACTION, PROCESS, STATE, and RELATION, but the categories have been subdivided into finer subcategories. Thus there are 13 subcategories for SUBSTANCE, 16 for ACTION, and 2 for PROCESS. The high level categories are convenient for implementing the grammar patterns because it is simpler to express the interaction patterns using the high level categories (i.e., SUBSTANCE ACTION$_{vp}$ SUBSTANCE covers many of the interactions) and therefore less grammar patterns are needed. However, these coarse patterns are too permissive because they allow invalid combinations. Therefore the subclasses, which are more specific, are used to constrain the arguments of the action verbs. Interestingly, these results are identical to the results found by Harris et al. [37] in the work on the immunology sublanguage. In order to illustrate this phenomenon, we note that the pattern CELL ACTIVATE$_{vp}$ GENE is not valid while the pattern CELL

Table 5
Co-occurrence patterns in the biomolecular sublanguage illustrated with example and target output forms

Category	Basic patterns	Examples	Target form
Action	Substance + action$_{vp\|ved}$ + substance	*Fyn activates Cbl*	[action,activate,[protein,Fyn],[protein,Cbl]]
	Substance + be + action$_{ven}$ + by + substance	*Cbl was activated by Fyn*	
	Substance + action$_n$ + of + substance	*Fyn activation of Cbl*	
	Action$_n$ + of + substance + by + substance	*Activation of Cbl by Fyn*	
	Substance + action$_{ved}$ + by + substance	*Cbl activated by Fyn*	
	Substance + action$_{vor}$ + substance	*Cbl activator Fyn*	
	Substance + action$_{vor}$ + of + substance	*Fyn activator of Cbl*	
	Action$_n$ + of + substance + and + substance	*The association of Cbl and Fyn*	[action,attach,[protein,Cbl],[protein,Fyn]]
	Substance + action$_n$ + with + substance	*Fyn association with Cbl*	
	Substance + dash + substance + action$_n$	*Fyn–Cbl association*	
	Action$_n$ + of + substance	*Transcription of Il-2 gene*	[action,transcribe,[gene,il-2]]
	Substance + action$_n$	*Il-2 gene transcription*	
Substance	Substance	*Bcl-2*	[protein,Bcl-2]
	State + substance	*Active Bcl-2*	[protein,Bcl-2,[state,active]]
	Substance + which/that + actionv$_{vp\|ved}$ + substance/process	*Bcl-2 that promotes cell death*	[protein,Bcl-2,[action,promote,[protein,Bcl-2],[process,apoptosis]]]
Process	Process	*Mitosis*	[process,mitosis]
	Substance + pathway	*Ras pathway*	[pathway,Ras]
Relation	Substance + conj + substance	*Bcl-2 and Bad*	[relation,and,[protein,Bcl-2],[protein,BAD]]

EXPRESS$_{vp}$ GENE is. Thus, in our work, we have found that a fine-grained subclassification of the action verbs, biomolecular substances and entities is consistent with the ontology specified by Rzhetsky et al. [58] and is useful to eliminate invalid patterns.

Table 5 shows some simplified patterns for the domain. Notice that there are many different ways of expressing the same information. In Table 5, there is a set of seven different action patterns that have the same target form (e.g., *Fyn activates Cbl*, *Cbl was activated by Fyn*, *Fyn activation of Cbl*, *Activation of Cbl by Fyn*, *Cbl activated by Fyn*, *Cbl activator Fyn*, *Fyn activator of Cbl*), which consists of an interaction **activate** which has two arguments: an agent **Fyn** and a target **Cbl**. The representation for the information is the following:

[**action**, **activate**, [**protein**, **Fyn**], [**protein**, **Cbl**]]

The target form is similar to that associated with the clinical domain. Table 6 provides examples of target forms for seven sentences taken from the literature, which are typical in that they are complex and nested. The target form is similar to that of MedLEE because it is also represented in the form of frame-based lists, where the first element of the list represents the type of information (e.g., **action**) and the second denotes the value (e.g., **activate**). However, the interpretation of the remaining elements is dependent on the frame being represented. In action and relation frames, the next two elements are generally ordered arguments of the action or relation, although some actions, such as **transcribe** have only one argument. In the above example, the first frame, [**protein**, **Fyn**], is the agent of the action whereas the second frame, [**protein**, **Cbl**], is the target. In that example there are no modifiers of the action, but if there were they would follow the arguments. For example, in the last sentence in Table 6, the action **phosphorylate** has two arguments followed by a negation modifier representing negation in the sentence *Inactive akt failed to phosphorylate BAD*. Similarly, in the third sentence, **transcribe** has only one argument, which is followed by an action **promote** which modifies the *transcription* action.

In frames associated with entities or states, the elements that follow the value element represent modifiers. For example, in Table 5, the target form for *active BCL-2* represents an entity and therefore is [**protein**,**BCL**,[**state**,**active**]] because the entity is a protein, which is *BCL-2*, and is modified by the state *active*. A more complex modifier of a substance can be a relative clause that is an action (e.g., *Phosphorylated Cbl coprecipitated with CrkL*, *which was constitutively associated with the GNRF C3G*). The second sentence in Table 6 shows the target form for a sentence containing a relative clause. It is quite complex, but not atypical, because it also contains a conjunction as well as nested interactions.

In Table 5, the patterns incorporate both semantic and syntactic constraints. For example, the first pattern requires an action verb that is a present tense verb, whereas the third pattern requires an action verb in the nominal form. More complex patterns that occur in the sublanguage are obtained by replacing SUBSTANCE in many of the patterns with a nominalized action pattern, thereby providing for nesting of interactions. For example, the complex nested relation (SUBSTANCE ACTION$_N$ OF SUBSTANCE) ACTION$_{VP}$ (ACTION$_N$ OF SUBSTANCE TO SUBSTANCE) is formed by substituting both the subject and object of the main action verb, as in *Akt*

Table 6
Examples of sentences with nested interactions and corresponding output representational forms

Interleukin-3-induced phosphorylation of BAD through the protein kinase akt [action,activate,[protein,interleukin-3],[action,phosphorylate,[protein, kinase akt],[protein,bad]]
The adapter protein crkl was associated with both phosphorylated cbl and the guanidine nucleotide-releasing factor c3g, which catalyzes guanosine triphosphate (gtp) exchange on rap1 [action,attach,[protein,crkl],[relation,and,[protein,cbl, [state,phosphorylated]],[protein,guanidine nucleotide-releasing factor c3g,[action, activate,[protein,guanidine nucleotide-releasing factor c3g],[smallmolecule,guanosine triphosphate]]]]
Activated rap1 functions as a negative regulator of tcr and cd28-mediated il-2 transcription [action,inactivate,[protein,rap1,[state,active]],[action,transcribe, [gene,gene encoding interleukin-2],[action,promote,[relation,and,[complex,tcr], [protein,cd28]],[action,transcribe,[gene encoding interleukin-2]]]]]
Inhibition of 4 e-bp1 phosphorylation enhanced 4 e-bp1 binding to eif-4e [action,promote,[action,inactivate,x,[action, phosphorylate,x,[protein,4 e-bp1]]],[action,attach,[protein,4 e-bp1],[protein,eif-4e]]]
Phosphorylation of catenin by glycogen-synthase kinase-3 induces the degradation of catenin by the ubiquitin-proteasome pathway [action,activate,[action,phosphorylate,[protein,glycogen-synthase kinase-3],[protein,catenin]],[action,degrade,[pathway, ubiquitin-proteasome],[protein,catenin]]]
BAD phosphorylation induced by Akt was not inhibited by wortmannin [action,inactivate,[protein,wortamannin],[action,activate,[protein,akt], [action,phosphorylate,x,[protein,bad]]],[certainty,no]]
Inactive akt failed to phosphorylate bad [action,phosphorylate,[protein,akt,[state,inactive]],[protein,bad],[certainty,no]]

phosphorylation of BAD precludes the binding of BAD to Bcl-xL. In this case the arguments of the verb *precludes* are also interactions. Further examples of nested interactions are shown in Table 6, where almost all the target forms demonstrate nesting.

6. Comparison of biomedical and biomolecular sublanguages

Since the sublanguage of a particular science reflects the underlying information of the science, it is not surprising that the clinical and biomolecular sublanguages are substantially different but also have a number of interesting similarities. While this paper focuses only on comparison between these particular two sublanguages, a more thorough analysis of sublanguage differences and commonalities was performed by Harris [28], who provides a theoretical basis for comparison through the notion of "prior science." In addition, Sager et al. [34] compares the patient care and pharmacology sublanguages, and also discusses common characteristics of the various science subfields.

The clinical sublanguage primarily expresses descriptions of entities and events associated with the patient state, whereas the biomolecular sublanguage expresses descriptions of events associated with biomolecular substances and their interactions. Some of the entities and modifiers in both domains overlap. Both domains have modifiers relating to evidential, change, quantitative, degree, and body location (referred to as tissue in the biomolecular domain) types of information. For example, in the clinical domain, sentences, such as, *pneumonia not ruled out* and *slight improvement in pneumonia* contain evidential (e.g., *not ruled out*) and change information (e.g., *improvement in*) that modify *pneumonia*, and degree information (e.g., *slight*) that modifies the rate of change (e.g., *improvement*). In the biomolecular domain, we also find evidential modifiers, such as *these results strongly suggest* as in *these results strongly suggest that constitutive activation of the PI3K/AKT pathway plays an essential role in v-Crk-induced transformation of CEF*, and change and degree modifiers, such as *significant increase* in the sentence *active Akt induced a significant increase in BAD phosphorylation*. Additionally, there is overlap in the subject matter because both languages are concerned with tissues, diseases, cells, and molecular components, such as genes and other types of disease markers, and therefore the grammars of the two sublanguages share these informational categories. For example, disease events and biomolecular interactions occur in both. However, disease events occur more frequently in the clinical domain whereas interactions occur more frequently in the biomolecular domain, reflecting the primary concerns of the respective domains. Disease events occur in the biomolecular domain in association with biomolecular interactions but there is little emphasis on their description. Biomolecular events occur in the clinical domain because they are related to testing for the presence of biological markers.

Typically, in clinical reports, the descriptions of diseases, symptoms, diagnostic procedures, and treatments are quite detailed, and often have many different types of modifiers associated with time, change, severity, body location, descriptive, and certainty types of information because accurate descriptions of these events are critical. Additionally, modifiers, which are secondary information, are themselves modified less frequently. When biomolecular substances and interactions occur in clinical reports, they primarily occur in pathology reports and denote findings (i.e., expression levels) of tests associated with biomolecular markers. Interactions, such as *neoplastic B cells do not express CD11C*, are found in text of both sublanguages. However, in pathology reports, the interaction typically is a measurement denoting the level of expression, and the types of interaction modifiers are limited as they mainly refer to negation and degree types of information. In the biomolecular domain, the situation is the opposite. Biomolecular interactions and relations are quite complex and highly nested but disease information is straightforward and occurs with few modifiers.

In both domains, the semantic relationships associated with the overlapping semantic categories are also quite different, reflecting the different types of relationships. For example, in the clinical domain, diseases are primarily associated with procedures (*V-Q scan positive for pulmonary embolism*), treatments (*on Bactrim for urinary tract infection*), and patient information occurs with temporal, severity, and body location types of modifiers (*chronic pulmonary embolism diagnosed in 1994*). In the biomolecular domain, diseases are associated with genomic variations (*alterations of the PPP2R1B gene were found in human lung and colon cancer*), and molecular functional information, which sometimes refers to a particular type of tissue/body location (e.g., *we evaluated CAR expression in prostate carcinoma*, *RSG16 is abundantly expressed in the retina*). In this domain, biomolecular substances and interactions frequently have modifiers, but tissues and diseases generally do not.

A major difference between the two domains is the complexity of the entities and relations. In the medical domain, the information primarily is descriptive of the patient's condition. Thus, the primary concepts (i.e., disease, procedure, medication, vital sign, symptom, and body location) are mostly nouns, and, excluding quantitative information, the modifiers are also generally adjectives or nouns. The relations among the classes can be divided into two types: simple and complex. A simple relation consists of a single finding and associated modifiers. In simple relations, verbs are frequently omitted (as well as the subject nouns) because they are expected and therefore have low information content (e.g., in *fever and headache*, the phrase *patient had* was omitted because it is expected); if verbs exist, they are used to connect findings to modifiers (e.g., *heart was enlarged*, *blood pressure was high*, *pulse measured 70 bpm*). A complex relation connects several findings, or connects findings to procedures and/or treatments using connective operators that are usually conjunctions (e.g., *and*, *with*), prepositions and verbs associated with causality (*due to*, *led to*), modality (*suggests*, *including*), and time (*status post*, *after*). In the biomolecular domain, the primary information concerns descriptions of biomolecular pathways consisting of complex interactions and other relations. The primary relations associated with the pathways are expressed using verbs because they denote interactions between substances (e.g., *p53 binds to il2*). Frequently the verbs are expressed in the corresponding nominal or noun form (e.g., *activation*) to allow for nesting. Since a pathway itself is complex and consists of sequences of interactions, the language expresses the sequences using complex and highly nested relations. Thus, an argument of an interaction can be another interaction and so forth, in which case the interaction that is an argument generally occurs in the sentence in the nominal form. For example, the fourth sentence in Table 6 (*Inhibition of 4 e-bp1 phosphorylation enhanced 4 e-bp1 binding to eif-4e*) illustrates a sentence where both arguments of the main verb *enhanced* are interactions. The subject of *enhanced* is an interaction *inhibition*, which is in the nominal form. It also has a nested interaction *4 e-bp1 phosphorylation*, which represents an additional level of nesting. Frequently, an additional level of nesting is expressed as a past participle modifying an interaction in the nominal form. For example, in the third sentence in Table 6 (*Activated rap1 functions as a negative regulator of tcr and cd28-mediated il-2 transcription*) *tcr and cd28-mediated* modify *transcription* which is an argument of *functions as a negative regulator*.

7. Conclusions

Following the sublanguage theory of Zellig Harris, we have delineated two specialized sublanguages, which are quite different from each other but have some overlapping components. A general English grammar would have restricted us to specifying only the syntactic structure of English, but by using the sublanguage approach, we were able to delineate the grammatical structure of each specialized language and intersperse syntactic information with the informational structure and content of the language. One sublanguage concerns the clinical domain, which is expressed in patient reports, which is descriptive in nature, and which is dominated by nouns and adjectives because the main subject matter consists of clinical findings, treatments, and procedures, which are expressed primarily as nouns. The second sublanguage is concerned with the biomolecular literature, which contains complex relations between biological substances, and which is dominated by relations based on verbs. Specification of the sublanguages was accomplished by establishing semantic categories for the entities and relations in the domain, specifying semantic and syntactic co-occurrence patterns, and specifying target forms for each of the patterns. The two grammars were implemented and incorporated into operational NLP systems, called MedLEE and GENIES, which both share a common processing engine. The common underlying theory of sublanguage made it possible for the two systems to share the same engine by changing only the sublanguage grammars. However, establishment of a sublanguage grammar is difficult and we accomplished it using manual analysis of sample corpora of the two domains. Future work will involve developing machine learning techniques to help automate or semi-automate the process of discovering new co-occurrence patterns.

Acknowledgments

We owe our thanks to Drs. Stephen Johnson and Pierre Zweigenbaum, who contributed to this paper by their valuable comments on an earlier draft. This work was supported in part by Grants LM06274 and GN61372-01A2 from the National Institutes of Health, Grants EIA-0121687 and IIS-98-17434 from the National Science Foundation, and a grant from the Columbia Center for Advanced Technology supported by the New York State Science and Technology Foundation.

References

[1] Fiszman M, Haug PJ. Using medical language processing to support real-time evaluation of guidelines. Proc AMIA Symp 2000:235–9.

[2] Friedman C, Knirsch CA, Shagina L, Hripcsak G. Automating a severity score guideline for community-acquired pneumonia employing medical language processing of discharge summaries. Proc AMIA Symp 1999:256–60.

[3] Gundersen ML, Haug PJ, Pryor TA, van Bree R, Koehler S, Bauer K, et al. Development and evaluation of a computerized admission diagnoses encoding system. Comput Biomed Res 1996;29:351–72.

[4] Blanquet A, Zweigenbaum P. A lexical method for assisted extraction and coding of ICD-10 diagnoses from free text patient discharge summaries. Proc AMIA Symp 99 Nov 6;1999:1029.

[5] Lussier Y, Shagina L, Friedman C. Automating SNOMED Coding using medical language understanding: a feasibility study.

In: Baaken S, editor. Proc 2001 AMIA. Phila: Hanley & Belfus; 2001. p. 418–22.

[6] Henize TD, Morris WC, Warner Jr HR, Morsch AEW, Sheffer RE, Jennings MA, et al. Assessing the accuracy of an automated coded system in emergency medicine. Proc 2000 AMIA Symp 2000:595–9.

[7] Doddi S, Marathe A, Ravi SS, Torney DC. Discovery of association rules in medical data. Med Inform Internet Med 2001;26(1):25–33.

[8] Hripcsak G, Austin J, Alderson PO, Friedman C. Use of natural language processing to translate clinical information from a database of 889,921 chest radiographic reports. Radiology 2002;224(1):157–63.

[9] Wilcox A, Hripcsak G. Medical text representations for inductive learning. Proc AMIA Symp 2000:923–7.

[10] Lovis C, Chapko MK, Martin DP, Payne TH, Baud RH, Hoey PJ, et al. Evaluation of a command-line parser-based order entry pathway for the Department of Veterans Affairs electronic patient record. J Am Med Inform Assoc 2001;8(5):486–98.

[11] Hersh W, Mailhot M, Arnott-Smith C, Lowe H. Selective automated indexing of findings and diagnoses in radiology reports. J Biomed Inform 2001;34(4):262–73.

[12] Chu S, Cesnik B. Knowledge representation and retrieval using conceptual graphs and free text document self-organisation techniques. Int J Med Inf 2001;62(2-3):121–33.

[13] Teufel S, Hatzivassiloglou V, Teufel S, McKeown KR, Jordan DA, Dunn KM, et al. Personalizing retrieval of journal articles for patient care. Proc AMIA Symp 2001:696–700.

[14] Aronson AR. Effective mapping of biomedical text to the UMLS metathesaurus: the MetaMap program. In: Bakken S, editor. 2001 AMIA Symposium. Hanley & Belfus; 2001. p. 17–21.

[15] Fukuda K, Tsunoda T, Tamura A, Takagi T. Information extraction: identifying protein names from biological papers. Proceedings of the Pacific Symposium on Biocomputing '98, Hawaii; 1998. p. 707–18.

[16] Jenssen T, Vinterbo SA. A set-covering approach to specific search for literature about human genes. In: Overhage M, editor. Proc 2000 AMIA Symp. 2000. p. 384–8.

[17] Krauthammer M, Rzhetsky A, Morozov P, Friedman C. Using BLAST for identifying gene and protein names in journal articles. GENE 2000;259(1/2):245–52.

[18] Tanabe L, Wilbur WJ. Tagging gene and protein names in biomedical text. Bioinformatics 2002;18(8):1124–32.

[19] Sekimizu T, Park HS, Tsujii J. Identifying the interaction between genes and gene products based on frequently seen verbs in medline abstracts. Nucleic Acids Res 1998;9:62–71.

[20] Humphreys K, Demetrion G, Gaizauskas R. Two applications of information extraction to biological science journal articles: enzyme interactions and protein structures. In: Proceedings of the 2000 Pacific Symposium on Biocomputing. 2000. p. 505–16.

[21] Blaschke C, Andrade MA, Ouzounis C, Valencia A. Automatic extraction of biological information from scientific text: protein–protein interactions. In: Proceedings International Conference on Intelligent Systems for Molecular Biology, Heidelberg. 1999. p. 60–7.

[22] Yakushiji A, Tateisi Y, Miyao Y, Tsujii J. Event extraction from biomedical papers using a full parser. Pac Symp Biocomput 2001:408–19.

[23] Thomas J, Milward D, Ouzounis C, Pulman S, Carroll M. Automatic extraction of protein interactions from scientific abstracts. Pac Symp Biocomput 2000:541–52.

[24] Rindflesch TC, Tanabe L, Weinstein JN, Hunter L. EDGAR: extraction of drugs, genes and relations from the biomedical literature. Pac Symp Biocomput 2000:517–28.

[25] Park JC, Kim HS, Kim JJ. Bidirectional incremental parsing for automatic pathway identification with combinatory categorial grammar. Pac Symp Biocomput 2001:396–407.

[26] Pustejovsky J, Castano J, Zhang J, Kotecki M, Cochran B. Robust relational parsing over biomedical literature: extracting inhibit relations. Pac Symp Biocomput 2002:362–73.

[27] Harris Z. A grammar of english on mathematical principles. New York: Wiley; 1982.

[28] Harris Z. A theory of language and information: a mathematical approach. Oxford: Clarendon Press; 1991.

[29] Grishman R, Kittredge R. Analyzing language in restricted domains: sublanguage description and processing. Hillsdale, NJ: Lawrence Erlbaum; 1986.

[30] Friedman C. Sublanguage text processing—application to medical narrative. In: Grishman R, Kittredge R, editors. Analyzing language in restricted domains. Hillsdale, NJ: Lawrence Erlbaum; 1986. p. 85–102.

[31] Fitzpatrick E, Bachenko J, Hindle D. Status of telegraphic sublanguages. In: Grishman R, Kittredge R, editors. Analyzing language in restricted domains: sublanguage description and processing. Hillsdale, NJ: Lawrence Erlbaum; 1986. p. 39–51.

[32] Marsh E. Semantic patterns in sublanguages. In: Grishman R, Kittredge R, editors. Analyzing language in restricted domains: sublanguage description and processing. Hillsdale, NJ: Lawrence Erlbaum; 1986. p. 103–27.

[33] Montgomery CA, Glover BC. Reporting and analysis of space events. In: Grishman R, Kittredge R, editors. Analyzing language in restricted domains: sublanguage description and processing. Hillsdale, NJ: Lawrence Erlbaum; 1986. p. 129–61.

[34] Sager N, Friedman C, Lyman M, et al. Medical language processing: computer management of narrative data. Reading, MA: Addison-Wesley; 1987.

[35] Kittredge R, Lehrberger J, editors. Sublanguage: studies of language in restricted semantic domains. Berlin: deGruyter; 1982.

[36] Sager N, Lyman M, Buchnall C, Nhan NT, TIck LJ. Natural language processing and the representation of clinical data. JAMIA 1994;1(2):142–60.

[37] Harris Z, Gottfried M, Ryckman T, Mattick P, Daladier A, Harris TN, et al. The form of information in science: analysis of an immunology sublanguage. Dordrecht: Kluwer Academic; 1989.

[38] Hirschman L, Grishman R. Grammatically-based automatic word class formation. Inform Proc Manag 1975;11:39–57.

[39] Friedman C, Huff SM, Hersh WR, Pattison-Gordon E, Cimino JJ. The canon group's effort: working toward a merged model. JAMIA 1995;2(1):4–18.

[40] Chute C, Elkin PL. A clinically deried terminology: qualification to reduction. In: Masys DR, editor. Proceedings of the 1997 AMIA Fall Annual Symposium. Phila: Hanley & Belfus; 1997. p. 570–4.

[41] Elkin PL, Tuttle MS, Keck K, Campbell K, Atkin G, Chute C. The role of compositionality in standardized problem list generation. In: Cesnik B, Safran C, Degoulet P, editors. Proceedings of MEDINFO 98. Amsterdam: IOS Press; 1998. p. 660–4.

[42] McCray AT, Browne AC. Discovery of modifiers in a terminology data set. In: Chute C, editor. Proceedings AMIA 98 Annual Symposium. Phila: Hanley & Belfus; 1998. p. 780–4.

[43] Manning CD, Schutze H. Foundations of statistical natural language processing. Cambridge: MIT Press; 1999.

[44] Lindberg D, Humphreys B, McCray AT. The unified medical language system. Meth Inform Med 1993;32:281–91.

[45] Cote RA, Rothwell DJ, Palotay JL, Beckett RS, Brochu L, editors. The systematised nomenclature of humans and veterinary medicine: SNOMED International. Northfield: College of American Pathologists; 1993.

[46] Department of Health and Human Services. International classification of diseases. Washington, 1990.

[47] Medical Subject Headings. http://www.nlm.gov/mesh: 2002.

[48] Ashburner M, Ball CA, Blake JA, Botstein D, Butler H, Cherry JM, et al. Gene ontology: tool for the unification of biology. Nat Genet 2000;25:25–9.

[49] Benson DA, Boguski MS, Lipman DJ, Ostell J, Francis BF. GenBank. Nucleic Acids Res 1998;26(1):1–7.

[50] Bairoch A, Apweiler R. The SWISS-PROT protein sequence database: its relevance to human molecular medical research. J Mol Med 1997;75(5):312–6.

[51] Pruitt KD, Maglott DR. RefSeq and LocusLink: NCBI gene-centered resources. Nucleic Acids Res 2001;29(1): 137–40.

[52] Friedman C, Alderson PO, Austin J, Cimino JJ, Johnson SB. A general natural language text processor for clinical radiology. JAMIA 1994;1(2):161–74.

[53] Sager N. Natural language processing: a computer grammar of english and its applications. Reading, MA: Addison-Wesley; 1981.

[54] Hirschman L, Palmer M, Dowding J, Dahl D, Linebarger M, Passonneau R, et al. The PUNDIT NLP system. AI Systems in Government Conference 89 March 27. Computer Society of IEEE, 1989.

[55] Palmer M, Passonneau R, Weir C, Finin T. The KERNEL text understanding system. In: Pereira FCN, Grosz B, editors. Natural language processing. Cambridge: MIT Press; 1994.

[56] Friedman C, Hripcsak G, DuMouchel W, Johnson SB, Clayton PD. Natural language processing in an operational clinical information system. Nat Lang Eng 1995;1(1):83–108.

[57] Friedman C, Kra P, Yu H, Krauthammer M, Rzhetsky A. GENIES: a natural-language processing system for the extraction of molecular pathways from journal articles. Bioinformatics 2001;17(Suppl 1):S74–82.

[58] Rzhetsky A, Koike T, Kalachikov S, Gomez SM, Krauthammer M, Kaplan S, Kra P. A knowledge model for analysis and simulation of regulatory networks. Bioimformatics 2000;16(12): 1120–8.